December 2018

The Architext of Biblion: Digital Echoes of Paul Otlet

Arthur Perret
Université Bordeaux Montaigne, arthurperret@me.com

Introduction: The Architect of Babel

Paul Otlet (1868–1944), a well-known figure among document scholars, dedicated his life to an ideal: peace through knowledge—building a better society by improving access to information, in the hope of reducing ignorance and fear. And while he may be regarded by some simply as an idealist, the architect of a dream, there is much to be said about his intellectual legacy.

In the latter part of his life, Otlet compiled decades of publications and personal documentation into his most important books: *Traité de documentation: Le livre sur le livre* (1934) and *Monde, essai d'universalisme* (1935). The *Traité* is widely considered to be the first manual of documentation.

Among many insights, it introduces the notion of *biblion*—a unit of information around which Otlet designs a framework for document theory (Robert, 2015). It is a fairly ambiguous term, referring to both media and meaning, the physical object (document or book) and the information it carries. This is because, in Otlet's view, information can take as many shapes as there are media to inscribe, far beyond the limited range of the book. A document is simply information recorded for transmission. He therefore uses a unit as a way to *handle* information on multiple levels: theoretically, because the idea of information beyond media is quite abstract; mechanically, as documents are transformed into index cards which are the units of a file system; mathematically, as information is encoded into a decimal classification.

The *Traité* contains a great number of fascinating statements, specifically in the way it echoes our own preoccupation with infobesity and misinformation. It had a role in the advent of documentation as a field of professional practice and research, with lasting impact on document theoreticians. It is also a daunting read: it contains 350,000 words, set in a two-column layout over 431 wide, quarto pages; it has only ever been reprinted twice, in facsimile editions (in 1989 and 2015); the style is very much encyclopedic, with an obsession for systematic description which has been described as at times tedious (Rayward, 2012). Thus the "Bible of documentation" metaphor comes to mind.

One of Otlet's projects was to build a World City, with information pathways closing the distance between men, and with knowledge as its beating heart. Though it never came to be, there are echoes of this Babelian enterprise in our digital age. Otlet's written work sheds some light on contemporary issues related to information; it also contributes to an epistemology of information science rooted in document theory.

In this paper, we focus on the *Traité* itself, specifically the way it can illustrate an intellectual lineage between the analog and digital environments, both conceptually and empirically.

From Biblion to Architext

In section 243 of his *Traité*, Otlet describes various "substitutes of the book" which, because of the technological advances of his time, represent a growing body of new documents: discs, films, performances, objects used as evidence, and many more. This notion sketches a very open definition of the document, which was expanded even further by Suzanne Briet and Robert Pagès (Buckland, 2017), becoming almost overwhelming in its scope.

The categorization of these "substitutes" is made possible by the biblion: a concept which lays the foundation for an atomistic view of information. The word itself shares the ambiguity of "book" or "document" in the context of Otlet's writing, where they are polysemic, often substituted for one another, and can designate different things depending on which part of the *Traité* they appear in. He writes:

> Until an agreement be made on unified terminology, we will use indifferently the terms formed of the following four radicals, two Greek, two Latin, giving them by convention an equivalent meaning: first, *biblion*; second, *grapho* (gram grammata); third, *liber*; fourth, *documentum*. (Otlet, 1934, p. 12, translation mine)

Consequently, he defines biblion as:

1. "a generic term for all species" of documents (p. 43)
2. "the intellectual, abstract unit" of information (p. 43)
3. "writing and text" (p. 372), "writings" (p. 373)

Therefore, biblion means document but also the information carried by a document, regardless of its specific shape. With this concept, Otlet theorized how information could take a more flexible form, far beyond the book.

The biblion is closely tied to writing and could be regarded as meaning data, for it opens a path to conceiving texts as databases. Indeed, with computing, we are moving from a document paradigm to another, loosely defined as data-centric, which is often presented as entirely new approach. However, while digital objects do vary in shape, dimension and granularity, they simply raise the same issue as Otlet's substitutes, Briet's antelope or Pagès's gorilla—the need for a conceptual framework to tie them together while being coherent with practical implications.

By defining documents in such manner, Otlet foreshadowed a non-linear read/write system, hypertext, but we will use another term, which provides high-level description: architext. The concept originated in literary studies, where architextuality refers to texts as part of genres (Genette, 1992). The word carried over to information science, where it was interpreted as the architecture which marks out text and governs its enunciation (Jeanneret & Souchier, 1999). Using the word *text* to designate a literary object as a whole semantic field (Treharne, 2009), the architext can be seen as:

- everything which is not *text* but related to it
- a form of writing that expresses *text*

DOI: 10.35492/docam/5/2/6

This concept is especially relevant in a digital environment, as it helps us understand how computing implements the delegation of some architectural function to writing itself, and what we can derive from that.

At a simple level, the architext is the markup that allows text to be structured and rendered in a specific way: It is a way of encoding text, with instructions made of words and delimiters, such as the iconic </> tags found in all SGML-derived languages (e.g., XML or HTML). At a higher level, the architext enables hyperdocuments by expressing links between texts: from a single URI to entire programming libraries, hypertextuality connects different types of documents with various levels of granularity—all this through markup.

It should be noted that architext does not mean metadata. In their most simple form, they seem to overlap: a title and date at the top of a sheet of paper are metadata and their documentary functions do contribute to the expression of text (stabilizing information, allowing for quicker reference, constituting evidence). However, a digital architext is mostly made of structural components which carry no information at all: intrinsically meaningless elements used to apply formatting (such as *div* and *span* tags in HTML); layout instructions written in code (such as JavaScript); anchors allowing for navigation; etc. The common aspect and the very bones of it all are non-alphabetical characters, either borrowed from punctuation or invented along the developments of typography—a veritable *scripturation* (Laufer, 1986) which warrants dedicated research of its own.

This "hyperdocumentation" is at the core of the *Traité*'s most difficult excerpts, in which Otlet anticipates a paradigm we are now living in (the Internet), while also describing things we cannot readily grasp—sometimes verging on the paranormal. Leaving that last part aside, we will focus here on how this framework of concepts can be applied in a very practical approach.

An Experiment in Digital Hermeneutics

> The *Traité de documentation* contains two sections, unequal in size. The longest one is a systematic description of the book and the document… The shortest section is dedicated to bibliology and it is of the utmost importance for this field of study. (Estivals, 1987, p. 13)

This is one example of a comment on Otlet's *Traité* that we can come across when scanning the literature in search of useful companion pieces to the book itself. It makes three statements, respectively about structure, content and significance. They could be verified at a glance using the table of contents as well as more in-depth literature on bibliology (Estivals, 1993, pp. 30–65), and then be made clearer through selective reading of the *Traité*. This would be the classic, qualitative process of text analysis.

In this article, our goal is to illustrate the benefits of a quantitative approach. By cross-referencing simple structural information with text statistics and classification, we are able to reach a similar level of description. More importantly,

it brings up observations that could not be made before, allowing us to formulate hypotheses from a different angle. As such, we aim to highlight the heuristic potential of exploring text as data.

We devised a small experiment which relies on the architext–biblion tandem. The former enables the latter: markup allows us to extract the intellectual content inside a digital document, as well as create distinct units of information inside it. This opens new possibilities in terms of processing. The flexibility of digital text means we can test the heuristic potential and hermeneutical value of several text structures and representations (e.g., list, table, graph).

We chose two complementary approaches:

1. transcribe the table of contents of the *Traité* as tabular data, then build structural representations
2. encode the entire content as raw text, then apply standard corpus analysis techniques (lexicometry)

A combination of three documents were used: the 2015 facsimile, the full text from Wikisource (https://fr.wikisource.org/wiki/Traité_de_documentation), and the EPUB version exported from the full text. The corpus file was formatted for processing with Iramuteq, with variables encoding the six main sections of the book.[1] The table of contents was revised and extended manually to include six levels of depth from a partially automated extraction based on regular expressions, then processed with RAWGraphs.

Hierarchical Data Visualization and Lexicometry

Schematization is fundamental to Otlet's approach. In particular, his archives contain many representations of networks as well as radiant and arborescent structures. The visualization methods we applied to the structural data draws from this focus on circular and structural imagery.

The circular dendrogram is a hierarchical tree arranged in a circle. Here, each node represents an entry in the table of contents, with links corresponding to ancestry and filiation. The node at the center of the figure represents the book. Nodes are ordered clockwise according to the numbering of the book.

Figure 1 shows the first level of the hierarchy, with a node representing the book at the center, and each of the six main sections placed clockwise according to their number.

Going deeper into the table of contents, the dendrogram shows an uneven distribution of subsections across the book, with parts 1 and 2 displaying many more ramifications than part 0. At depth level six, the complexity of the structure is made quite apparent.

[1] 0. Fundamenta; 1. La Bibliologie ou Documentologie; 2. Le livre et le document; 3. Le livre et le document. Unités ou Ensembles; 4. Organisation rationnelle du Livre et du Document; 5. Synthèse bibliologique).

DOI: 10.35492/docam/5/2/6

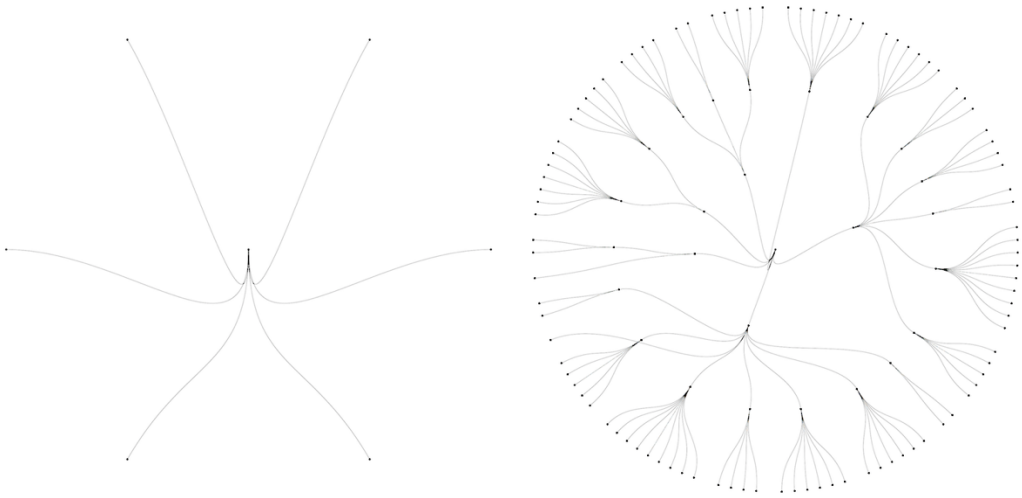

Figure 1. Circular dendrogram showing the six main sections, at depths of level 1 (left) and level 3

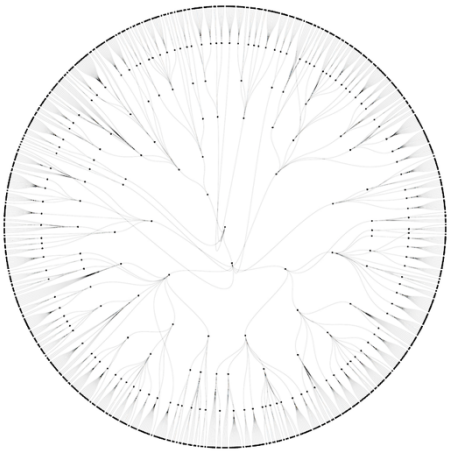

Figure 2. Circular dendrogram showing 6 levels of subsections

To gain a sense of the sections' relative proportions, we apply another method, the treemap (Figure 3). Here, each block represents a level-2 entry in the table of contents. Blocks are grouped by sections, with slightly larger spacing between groups to better distinguish the six sections. We then input the word count for each entry, therefore resizing the blocks to match their relative proportions. The treemap shows a striking difference in volume across sections, with part 2 (*Le livre et le document*) clearly representing the biggest segment of the book.

In order to use our hierarchical data in a meaningful way, we move on to an analysis of the full text. The first and most simple method we apply is a word cloud, which represents word frequency across the *Traité* (Figure 4).

The title of the book is *Traité de documentation* but its subtitle is *Le livre sur le livre*. Given how interchangeable the words "document" and "book" are in Otlet's writings, it could come off as a surprise that the latter dominates the numbers so clearly. It goes to show how important it is in Otlet's argumentation.

In essence, a word cloud suggests which ideas are at the core of a text, with further verifications required to make that claim with absolute certainty. A similarities analysis can give us a first glimpse at the lexical repartition, informing us of the relationships between the most frequent words in context.

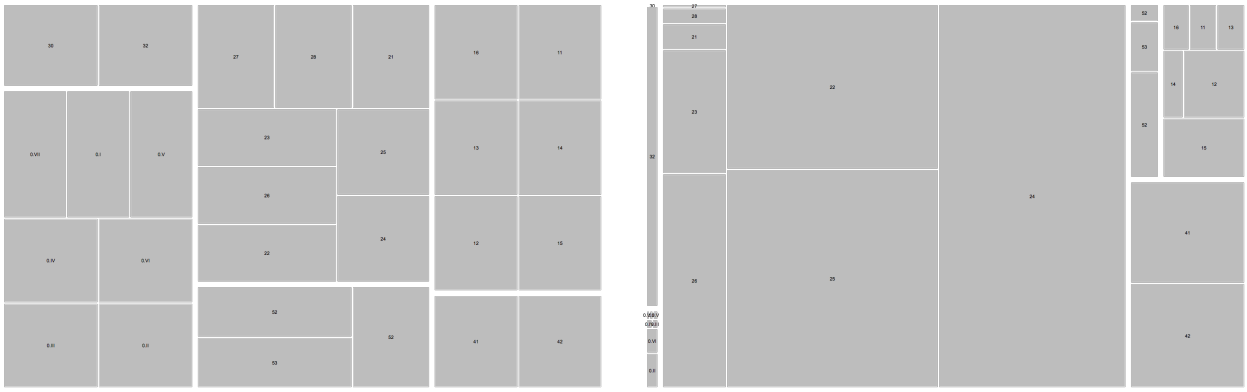

Figure 3. Treemap showing the main sections and subsections (right: proportional)

Figure 4. Most frequent words encountered in the book

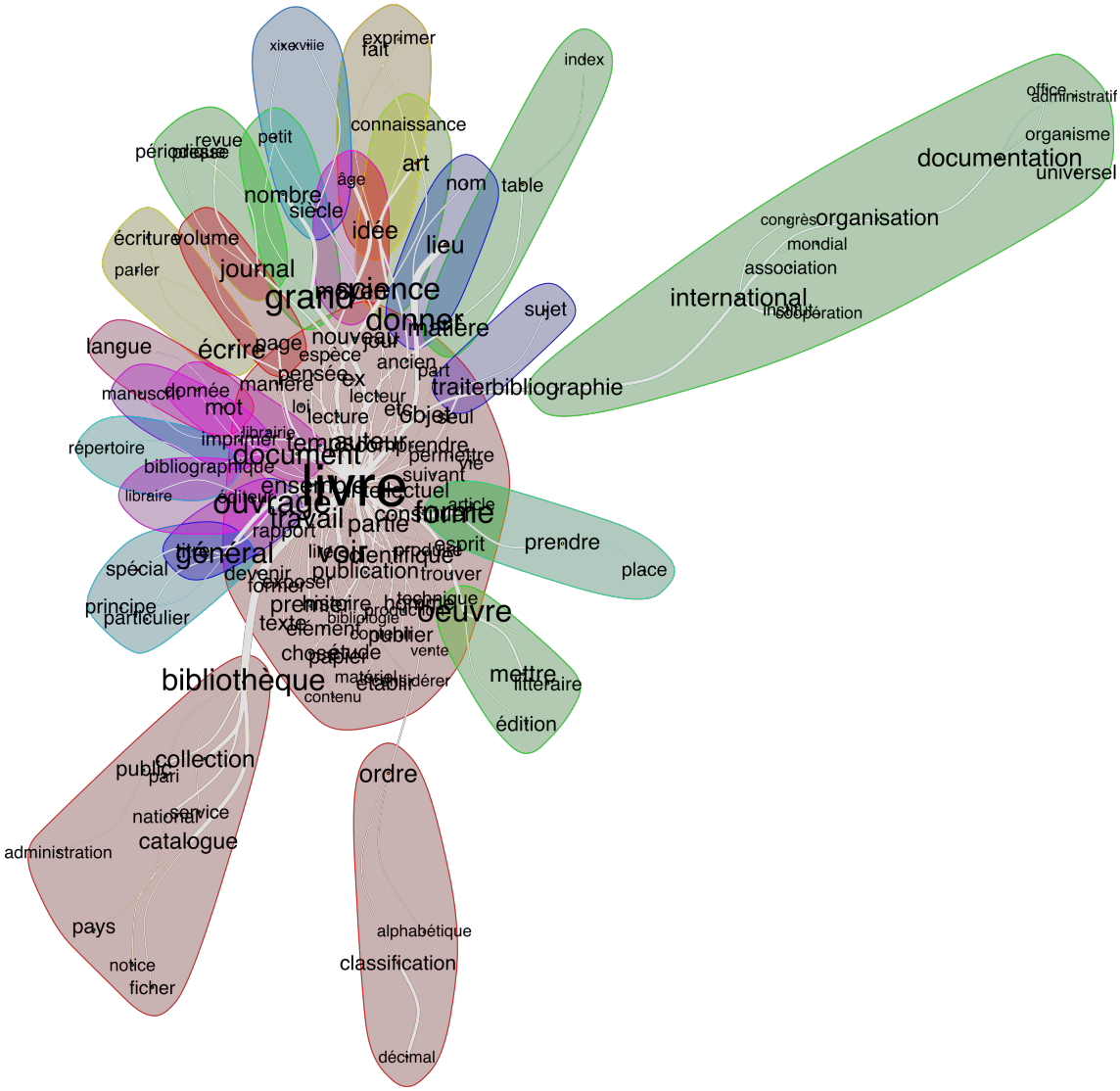

Figure 5. Similarity analysis showing relative homogeneity across the book

It is a somewhat difficult representation to work with. Readability and size are inversely proportional, which means that the surface of a work-in-progress is usually significantly bigger than that of the figure shared in a paper. Nevertheless, the flower-like distribution is a good indicator of homogeneity in a corpus; here, it confirms that the word “book” is not simply the most frequent word in the text but also the most central idea in it. “Documentation” stands out, as it not directly related to the word “book”: it is linked with the organizational aspects of Otlet’s work, with international cooperation appearing as a structuring parameter in the use and perhaps the definition of the word.

The bulk of the lexicometry depends on the classification and subsequent correspondence analysis. A global snapshot of the lexical profile is sufficient to glimpse the contents of the book: with 5 classes, we can distinguish the bibliographical description, the organization of knowledge and the matters of science. However, we wish for a more accurate profile, which is why we move on to a hierarchical descending classification (Reinert, 1983). We settle empirically for a setting which yields the most meaningful distribution, resulting in 12 lexical classes. Figure 6 shows the result; word size is not correlated to frequency but specificity.

Since the division of the *Traité* in parts was encoded as variables, we can plot them to obtain their lexical repartition. Figure 7 shows that, as far as lexical classes are concerned, there is a clear separation between two sets of book parts: [0, 1, 4, 5] and [2, 3].

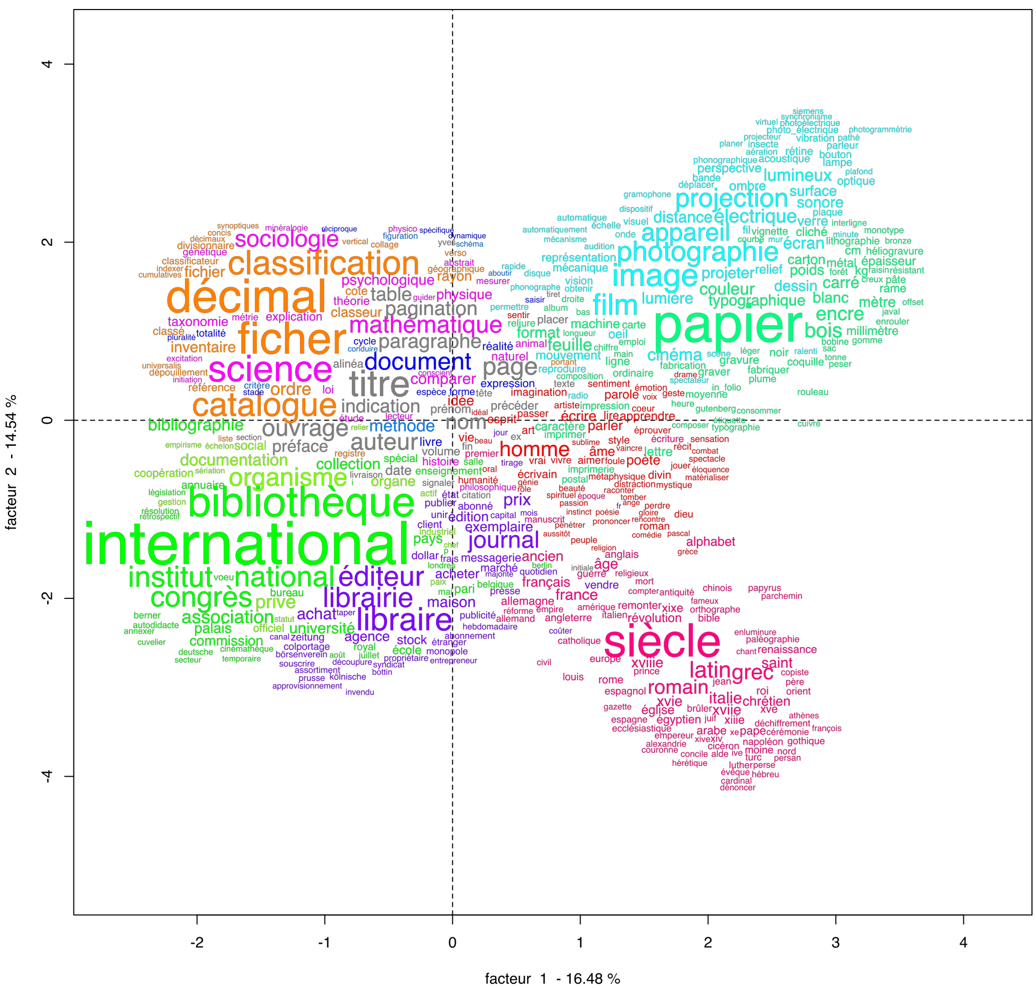

Figure 6. Proposed lexical classification

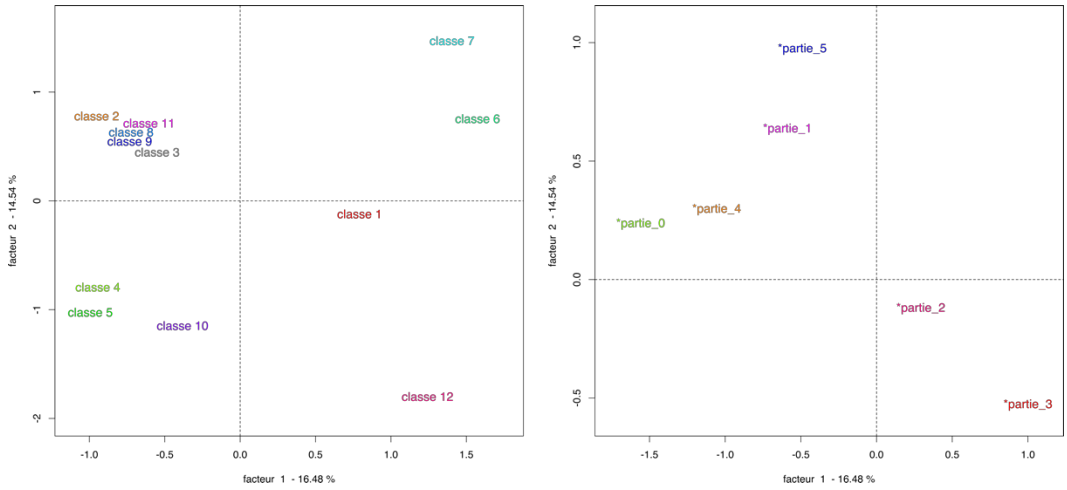

Figure 7. Lexical repartition of the book parts

How do we link parts and classes? This is where statistics are of great interest: since they are not readily available in a qualitative approach, they bring up interesting observations that may have come up much later otherwise, if at all. We look in particular at frequency, which is a simple count that can also be calculated relatively, and specificity, which results from a chi-square test.

Table 1 indicates whether the specificity of book parts to each class is positive (+) or negative (–). We judged the specificity score based on a significance criterion, aiming to highlight true positives: a low positive score in a short book part was not deemed significant and therefore treated as a negative. A brief but necessarily reductive description of each class' dominant aspects is provided, to help with the data's legibility.

The specificity score can also be used to look at smaller units of text, namely word forms, as seen in Table 2.

Discussion

Our results show indeed that the *Traité* is a two-legged, if somewhat lopsided, piece of work, with an overgrown bibliographic section bookended by shorter but dense epistemological work.

Figure 7 and Table 1 all but confirm this division. However, the data also underline the finer dynamics of the first set [0, 1, 4, 5]. Parts 0 and 5, being the introduction and conclusion, present their own variation on a common lexical profile; this reflects the necessary mix of vocabulary used in such context and is not surprising. Much more interesting is the difference between the other two, with part 1 seemingly containing most of the epistemological effort, while part 4 moves the need for a document science to its systematic application, with a sense of

Table 1. Book part specificity depending on class

class	part 0	part 1	part 2	part 3	part 4	part 5	description (suggested)
1	–	–	+	+	–	+	humanities & spirituality
2	+	–	–	–	+	–	documentation methods
3	–	–	+	–	–	–	bibliographical information
4	+	–	–	–	+	–	organization, society & politics
5	–	–	–	–	+	–	knowledge institutions & communities
6	–	–	+	–	–	–	material bibliography
7	–	–	+	–	–	–	media technologies
8	–	+	–	–	–	+	epistemology
9	+	+	–	–	+	+	document science
10	–	–	+	–	–	–	publishing & economy
11	–	+	–	–	–	+	sciences
12	–	–	+	+	–	–	history & historiography

Table 2. Word form specificity according to book part, with frequency

form	part 0	part 1	part 2	part 3	part 4	part 5	freq.
livre	–0.3	22.8	–9.2	–0.4	–9.5	20.4	2048
grand	–0.3	–1.3	1.7	0.3	–2.0	1.7	829
bibliothèque	1.5	–9.0	–0.8	–4.2	15.2	–6.8	781
science	0.4	65.6	–45.9	0.8	0.5	3.4	774
document	6.6	4.4	–24.1	–2.1	15.5	2.3	638

urgency brought by the technical, social and political challenges of Otlet's time. There is a common theme, but it is weaved differently.

This brings up the question of which thread was pulled. We know that in the following decades, scientific bibliology was almost abandoned, save for the occasional remembrance, while documentation thrived as a new area of practice. It calls to question whether the contents of part 4 were simply deemed more achievable by Otlet's readers, as opposed to the daunting prospects of inventing a new science, even though they were so closely linked. Perhaps a greater clarity of purpose played a part in consolidating documentation, as shown by the contributions of Suzanne Briet and her students (not least among them Robert Pagès). Bibliology, on the other hand, has remained a minor subject—although for reasons which are not limited to the *Traité de documentation*.

The data presented in Table 2 brings up another observation. The word frequency values for "book" and "document" are very high; they are at the heart of the *Traité*, as illustrated by the word cloud in Figure 4. Because of the sheer amount of times they occur, and taking into account the size of each book part, their low specificity to [2, 3] comes as a bit of a surprise. It is as if Otlet extracted the words from material bibliography and tied them irrevocably to a singular idea, blurring the lines between the terms. However, this ambiguity is not accidental: we have seen that he actually argues for the indifferent use of *biblion*, *grapho* or *gramma*, *liber* or *documentum* to form concepts until a consensus is reached.

Can we say that this consensus has indeed been reached? What about the importance that *data* has taken nowadays? Again, this can be tied to the question of Otlet's epistemological legacy. We know that the *Traité* belongs to a certain lineage, that it represents the culmination of a life's work for Otlet but also some of his colleagues and of course their predecessors working on bibliology; we also know how the book was received and the discreet influence it had in the following years. However, we know less about the extension of this lineage into the end of the 20th century and the beginning of the 21st. New approaches have been developed to adapt to a seemingly new information paradigm; the fate and relevance of Otlet's conceptual choices could be studied, perhaps with a mix of qualitative and quantitative methods.

Leaving these questions aside for another, more expansive study, there are two final considerations to be made.

Firstly, we now have many powerful tools that support different hermeneutical approaches to documents in general and text in particular. They sometimes yield quick results, in which case they should be used with twice as much caution, to avoid snowballing into absurd conclusions. As a general rule, these tools not only benefit from being articulated with a coherent theoretical framework, they require it to make any sort of significant observation, as small as it may be.

Here, we hope to have demonstrated the interest that lies in a science of writing that informs both concept and experiment. The goal was to show what

information quantitative methods bring to the table and how they feed back into a reflection on the text, its interpretation, its significance. Lexicometry is especially interesting for the study of theories: it provides data and representations for key concepts from a corpus, informing us on the correlations between structure and meaning.

Secondly, visual methods should not be seen as a simple means to an end, a technique used to produce a support for communication. They constitute a proper methodology as well, providing a way to test assumptions and explore sources. This is especially apparent when working with real-time rendering, which stimulates experimental approaches. Of course, this does not exclude the matter of output and exports, as the figures in this paper show. It simply means to reiterate that all forms of writing play a complex part in the way we think and work—something which Otlet probably had in mind when he included schematization in the constitutive elements of bibliology, the science of writing.

References

Buckland, M. K. (2017). Before the antelope: Robert Pagès on documents. *Proceedings from the Document Academy*, *4*(2), Article 6. Retrieved from http://ideaexchange.uakron.edu/docam/vol4/iss2/6

Estivals, R. (1987). *La bibliologie*. Paris: Presses Universitaires de France.

Estivals, R. (Ed.). (1993). *Les sciences de l'écrit*. Paris: Retz.

Genette, G. (1992). *The architext: An introduction*. Berkeley: University of California Press.

Jeanneret, Y., & Souchier, E. (1999). Pour une poétique de « l'écrit d'écran ». *Xoana, 6*, 97–107.

Laufer, R. (1986). L'énonciation typographique. *Communication et Langages, 1986*(68), 68–85.

Otlet, P. (1934). *Traité de documentation. Le livre sur le livre*. Brussels: Palais Mondial.

Rayward, B. W. (1975). *The universe of information: The work of Paul Otlet for documentation and international organisation*. Moscow: All-Union Institute for Scientific and Technical Information.

Rayward, W. B. (2012). Paul Otlet, an encounter. *Cahiers de La Documentation – Bladen Voor Documentatie*, *2*, 71.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Les Cahiers de L'Analyse des Données, 8*(2), 187–198.

Robert, P. (2015). Le biblion et les substituts du livre. Théorie et pratique du dépassement du livre chez Paul Otlet. *Communication et Langages*, *2015*(184), 3–23.

Treharne, E. (2009). The architextual editing of early English. *Poetica, 71*, 1–13.

DOI: 10.35492/docam/5/2/6